Software Requirements Specification for COMPSCI 4ZP6

RatBAT: Rat Behavioral Analysis Tool

Team 8

Brandon Carrasco

Daniel Locke

Jamie Wong

Inoday Yadav

Table of Contents

1 Document Information

1.1 Revision History

Date	Version	Notes
10/11/2024	0	Original draft of document
03/27/2025	1	Final version of document

1.2 Project Personnel

Name	Email	Role
Brandon Carrasco	carrascb@mcmaster.ca	Project Manager, Developer
Daniel Locke	locked3@mcmaster.ca	Developer
Jamie Wong	wongj171@mcmaster.ca	Developer
Inoday Yadav	yadavi1@mcmaster.ca	Developer
Anna Dvorkin-Gheva	dvorkin@mcmaster.ca	Supervisor
Henry Szechtman	szechtma@mcmaster.ca	Supervisor

1.3 Contributions

Name	Contributions
Brandon Carrasco	2, 4, 5, 6
Daniel Locke	2, 4, 5, 6
Jamie Wong	2, 5, 7
Inoday Yadav	2, 3, 7

2 Glossary

Term	Definition
Software	The entirety of the software, including the web platform, database, and Python package of summary measures.
System	The web platform, database, and related software of the project.
Interface	The graphical user interface (GUI) of the System.
FRDR	Federated Research Data Repository - A repository for Canadian research data operated by the Digital Research Alliance of Canada. *Source: FRDR*
OCD	Obsessive-Compulsive Disorder - A long-lasting disorder in which a person experiences uncontrollable and recurring thoughts (obsessions), engages in repetitive behaviours (compulsions), or both. *Source: National Institute of Mental Health*
Library	A Digital Library of Behavioural Performance in Standardized Conditions - A virtual library on the FRDR comprised of multiple datasets containing data related to experiments using an animal model of OCD. *Source: GigaScience, Volume 11, 2022*
Project	A collection of studies found in the library each with a common theme. *Source: GigaScience, Volume 11, 2022*
Study	An individual study corresponding to a specific dataset in the library. Each study consists of one or more experiments. *Source: GigaScience, Volume 11, 2022*
Experiment	A specific experiment within a study consisting of multiple related

	trials. *Source: GigaScience, Volume 11, 2022*
Trial	An individual test consisting of a rodent being placed on an open field for 55 minutes. For each trial the library contains at least 1 of 3 raw data objects: raw video, time series data, and trajectory plots *Source: GigaScience, Volume 11, 2022*
Raw Video	Raw video recordings of each trial. Video data is stored in .mpg files. *Source: GigaScience, Volume 11, 2022*
Time Series Data	A series of (x, y, t) spatial-time coordinates of each trial derived from the raw video. Each positional coordinate represents the center of mass of the rodent at that time. Data also includes values for a set of metadata variables pertaining to the trial. Time series data is stored in .csv files. *Source: GigaScience, Volume 11, 2022*
Trajectory Plots	Plots of animal trajectory over the course of each trial. Plots are derived from smoothed time series data and stored in .gif files. *Source: GigaScience, Volume 11, 2022*
Metadata Variables	A set of variables defined for each trial detailing specific conditions for that trial. Source: GigaScience, Volume 11, 2022
Animal Model	A non-human species used in biomedical research because it mimics aspects of a biological process or disease found in humans. For all projects in the library, researchers used the strain of Long-Evans Rats as test subjects. *Source: Animal Model (genome.gov)*
Quinpirole	A substance that acts as a dopamine D2/3 receptor agonist. *Source: PubMed Central 24406720*

Quinpirole Sensitization Rat Model (QSM)	An established pharmacological animal model of OCD based on chronic administration of quinpirole. The treatment induces compulsive-checking behaviour that is phenomenologically similar to human compulsive checking rituals. *Source: PubMed Central 27833539*
Open Field	Flat table containing zero or more small structures used for observing behaviour of rodents for experiments. *Source: GigaScience, Volume 11, 2022*
Key Locale	A place in the open field that the animal visited most often during a trial. Used as a point of reference for determining compulsive behaviour. *Source: Pubmed 29194070*
Summary Measure	Any measurement or information extracted from the spatial-time data gathered from trials. Examples include amounts and durations of activity of the animal during a trial and the frequency with which the animal returns to their key locale in a given interval of time.
MA	Moving Average Smoothing - A smoothing technique that reduces noise by setting data points to the average value of surrounding data points within a fixed sized window centred around that data point. *Source: Journal of Neuroscience Methods 133 (2004) 161-172*
LP	Local Polynomials Smoothing - A smoothing technique that reduces noise by fitting low-order polynomials over fixed size of the data and setting data points to points on these polynomials. *Source: Journal of Neuroscience Methods 133 (2004) 161-172*
LOWESS	Locally Weighted Scatterplot Smoothing - A smoothing technique that iteratively fits a series of polynomials to a dataset, then reduces the influence of outliers by weighting points based on their proximity to that polynomial. This process repeats until changes no longer occur. *Source: Journal of Neuroscience Methods 133 (2004) 161-172*

RRM	Repeated Running Median - A smoothing technique that reduces noise by iteratively setting fixed data points to the median of nearby data points within a window centred around that data point. *Source: Journal of Neuroscience Methods 133 (2004) 161-172*
SPSM	SEE Path Smoother - A smoothing technique that combines LOWESS and RRM to best capture all aspects of rodent movement. This technique will be the basis of our smoothing algorithm. *Source: Journal of Neuroscience Methods 133 (2004) 161-172*
GMM	Gaussian Mixture Model - A model used to represent a population with multiple distinct components each approximately normally distributed. In our case segments of lingering will be represented by a single gaussian while the remainder of the gaussian components will represent progression segments. *Source: Journal of Neuroscience Methods 96 (2000) 119-131*
EM	Expectation Maximization - An iterative algorithm that estimates the maximum likelihood parameters for a given number of components. In the case of EM on a GMM each component will be a gaussian distribution. *Source: Journal of Neuroscience Methods 96 (2000) 119-131*
SEE Path Segmentation	A segmentation technique that tests many iterations of the EM algorithm with different parameters to fit a GMM and then uses that gaussian mixture model to separate movements into progression and lingering. *Source: Journal of Neuroscience Methods 96 (2000) 119-131*
Progression	A class of rodent movement where the rodent is moving from one location to another. *Source: Journal of Neuroscience Methods 133 (2004) 161-172*
Lingering	A class of rodent movement where the rodent is only moving locally (moving around its current location rather than travelling to a

	different location). *Source: Journal of Neuroscience Methods 133 (2004) 161-172*
Arrest	A class of rodent movement where the rodent is immobile. *Source: Journal of Neuroscience Methods 133 (2004) 161-172*
EthoVision	A software produced by Dutch software company Noldus that is specifically designed to capture time series data from video of rodent movement. Version 3.0/3.1 of this technology was used to produce all time series data in the library. *Source: GigaScience, Volume 11, 2022*
Precision level noise	Source of noise in time series data which results from a rodent's centre of mass sitting on the border between two pixels and the EthoVision software inaccurately displaying oscillation between these points when a rodent is not in motion. *Source: Journal of Neuroscience Methods 133 (2004) 161-172*
Erratic system behaviour	Source of noise in time series data which results from unexpected behaviour displayed by the EthoVision software. *Source: Journal of Neuroscience Methods 133 (2004) 161-172*
Body wobble	Small body movements from a rodent that are not a part of its whole-body progression and which impact the EthoVision software's ability to measure the rodent's centre of gravity. *Source: Journal of Neuroscience Methods 133 (2004) 161-172*
Time to First Byte (TTFB)	The duration between a user sending a request to a server and the user's browser receiving the first byte of data from the server. *Source: NSF Public Access Repository 10192202*
Accuracy	The percentage of correctly classified labels after prediction.
Precision	The proportion of correctly classified positive labels to all predicted positive labels (true positives + false positives) after prediction.

Recall	The proportion of correctly classified positive labels to actual positive labels (true positives + false negatives) after prediction.
Precision-Recall Curve	The plot of recall vs. precision that illustrates the best "cut-off" points (defined by confusion matrices of different predictive models) to maximize the recall and/or precision.
MIT License	Open source software license.

3 Project Overview

3.1 Purpose

This project aims to leverage and expand upon the existing dataset generated from the Quinpirole Sensitization Rat Model of OCD experiments, which involved tracking rat movements in an open field and collecting x, y, t time-spatial data. Such experiments had run for several decades in the laboratory of Dr. Szechtman at McMaster University. Recently, the collected data had been annotated and deposited in a public repository, the FRDR, with the expectation that the data will be utilized by others in future research. However, though there exists a substantial amount of raw data, there is currently no easy and straightforward method to access, process, and analyse this stored data. The purpose of the present project is to develop a robust and accessible open platform for researchers to preprocess data, compute summary measures, analyse, and gather rat behaviour data from the FRDR repository. This platform is expected to be valuable for many projects, including creating a talking animal model of OCD by translating the x, y, t coordinate data of rat locomotion into an audio narrative.

3.2 Client and Stakeholders

Primary Stakeholders:

- **Our supervisors** are involved in the project by providing detailed feedback and outlining the specific functionalities they need from this platform. Their primary objective is to leverage the platform to create a Talking Animal Model of OCD, which will transform rat time-spatial data into an audio narrative.
- **Researchers** using or hoping to use data found in the library for their research and studies.

Supporting Stakeholders:

- **FRDR:** The repository hosting the raw data files.
- **RHPCS:** The organisation to permanently host the platform.
- **McMaster University:** Oversees RHPCS, who are hosting the platform.
- **Data Producers:** Authors and creators of the raw data.

- **Developers:** We, as the developers of these platforms and tools, are stakeholders in the quality of the work produced.

4 Functional Requirements

4.1 P0 Functional Requirements

1. The system must allow users to filter data by project, experiment, and metadata variables.
2. The system must allow users to download all structured time-spatial data that exists in the library.
3. The system must allow users to download links to all videos and path plots that exist in the library.
4. The system must temporarily store data downloaded from the FRDR in a local database.
5. The system must allow users to compute summary measures on temporarily stored pre-processed time series data.
6. The system must temporarily store computed summary measures.
7. The system must allow users to download any temporarily stored data to their local machine.
8. The system must allow users to apply the SPSM smoothing method to time series data.
9. The system must be able to partition smoothed data into episodes of progression and lingering using the SEE Path Segmentation method.
10. The system must temporarily store pre-processed time series data.
11. The system must store all metadata variables in a local relational database.

4.2 P1 Functional Requirements

1. The system must allow for users to modify the parameters of the SPSM smoothing algorithm before it is run.
2. The system must allow for users to modify the parameters of the SEE segmentation algorithm before it is run.
3. The system must store the preprocessing parameters used to compute stored data and link them to the stored data.
4. The system must return stored preprocessed data when the computed data already exists within the database.

4.3 P2 Functional Requirements

1. The system must allow users to input their own Python-based summary measure algorithms onto the website.
2. The system must store summary measures algorithms created by a user to be accessible for future use via that user's credentials.
3. The system must allow users to undo any processing steps performed to data.
4. The system must be able to compute summary measures that compare data across multiple trials.

5. The system must display any raw or pre-processed time series selected by the user in the form of path plots.

4.4 P3 Functional Requirements

1. The system must display raw video data on the website when selected by the user to preview.
2. The system must display trajectory plots on the website when selected by the user to preview.

5 Data and Metrics

5.1 Data Overview

Data Library: FRDR

This project is centred around the data library produced by the Szechtman Lab over the course of 2 decades of research. This data spans multiple studies involving the QSM, including testing a multitude of different independent variables. In total, this library is composed of 29 datasets corresponding to 29 studies which are spread out among 12 different projects. Each study involves one or more experiments which are each composed of multiple trials, totalling 43 experiments and nearly 20500 trials across the entire library.

Each trial is filmed producing the raw video, and then the raw time series data is extracted from the raw video using EthoVision software. While the EthoVision technology is largely accurate, there are a number of sources of noise in this time series data including body wobble, equipment quality, precision level noise, and erratic system behaviour. As a result, for summary measures to be computed accurately, a smoothing algorithm is needed to remove much of this noise. To provide further behavioural context necessary for summary measures to be computed, a segmentation of the time-spatial data is used, partitioning the data into episodes of lingering and progression movement.

The extracted time series data is also used to produce the trajectory plots. In addition, each time series data file is documented with values for an extensive list of metadata variables including an ID for the specific rodent used in the trial. These IDs correspond to detailed descriptions of each animal published in *GigaScience, Volume 11, 2022*, giving potential for extensive analysis and comparison of trials through the summary measure portion of this project. For each trial, one or more of these raw data documents exist in the library.

In total this library is composed of roughly 10 terabytes of raw data, all stored in the FRDR across multiple datasets.

5.2 Performance Metrics

Performance of the web platform and interface will be measured by web platform response time. Web platform response time will be measured in terms of Time to First Byte (TTFB) and server response time.

The ideal accuracy measure for the smoothing method is how closely it matches to the actual behaviour of the rodents in the raw video, but as evidenced by the noise produced when producing the time series data, it is infeasible to produce a testing set that is confirmed to exactly match the rodent's behaviour.

Instead we can measure the performance of our smoothing algorithm by comparing its performance to that of other algorithms on the same data. *Journal of Neuroscience Methods 133 (2004) 161-172* demonstrates how SPSM, the proposed ideal smoothing algorithm for rodent time series data, is shown to be accurate via comparison to known flaws in other algorithms. In cases with an inactive rodent (actual distance travelled near 0 over a segment of time), we will measure the accuracy of our algorithm by comparing the total distance travelled (caused by noise falsely suggesting movement by the animal) over a segment of inactivity to the total distance computed from the same segment smoothed with the MA smoothing algorithm with the aim to have a lower total distance than that produced by MA. On the other hand, in samples of high rodent activity, we will aim to have a higher distance travelled using our smoothing implementation than that produced using the MA algorithm. This metric is based on the above paper which demonstrates that the MA algorithm results in inaccurately low distances when a rodent is active and inaccurately high distances when a rodent is inactive.

An additional metric for the accuracy of the smoothing method will be the performance of the subsequent segmentation component (segmentation of smoothed time series data into episodes of lingering or progression). The quality of the smoothing algorithm can be measured by its ability to accurately segment the data, particularly at inflection points between a rodent's lingering and progression episodes. As this can be defined as a classification task (assigning data points to the lingering or progression class), classification metrics can be applied here. Some of the main measures that will be used are: accuracy, precision, recall, and precision-recall curves. As per *Journal of Neuroscience Methods 133 (2004) 161-172*, we will aim for better performance on these metrics using our smoothing algorithm than is achieved using the MA, LP, or LOWESS algorithms.

6 Non-Functional Requirements

6.1 Look and Feel Requirements

- The interface shall hide optional customization features under collapsible menus in order to provide a simple and user friendly display.
- The interface shall display time-series data in a lucid and manipulable format.
- The interface shall provide feedback to user actions within 1 second.

6.2 Usability and Humanity Requirements

- The interface shall be easily accessible to users of all technical skill levels.
- The system shall display clear error messages to the user whenever a function of the system fails due to user input.
- The system shall notify the user if one of the requested data types is unavailable for at least one of the trials they are attempting to download.

6.3 Performance Requirements

- The system shall not perform unnecessary computations on data.
- Data partitioning into lingering and progression episodes shall be performed with an accuracy of at least 90%.
- All algorithms used to compute summary measures shall be implemented in polynomial time.
- Data smoothing shall perform better than MA as defined in *(5.2)*.

6.4 Legal Requirements

- All interactions with the FRDR will be compliant with the FRDR terms of use.
- The software shall be open-source under the MIT license.
- The data accessible from the web platform shall be open source.

7 Constraints and Assumptions

7.1 Constraints

- **System Constraints**
  - The system must be hosted on a McMaster domain.
  - The system shall be built using only open source software and resources.
  - The system must be able to function given the limited storage on the host server.
- **Time Constraints**
  - As per the requirements of the course, the initial proof of concept must be ready by November 2024, and a fully functioning version should be completed by the end of March 2025.
- **Data Constraints**
  - Dependency on FRDR for data access and management capabilities.
  - Temporarily storing large amounts of time series data and processed information locally or on a server will be constrained by available disk space and memory.
  - For some trials in the library, not all raw forms of data are available which will limit the analysis that can be performed for certain experiments.

7.2 General Assumptions

- Researchers have a basic understanding of data preprocessing, analysis techniques, and how to work with time-spatial data.
- The FRDR repository will remain stable and accessible throughout the project duration.
- Users will have access to the internet throughout their usage of the platform.

7.3 Predicted Risks

- Noise in the raw data from the QSM experiments could negatively impact the quality of the summary measures.
- Unforeseen changes in FRDR API could cause delays.
- Unexpected technical issues affecting FRDR could impact the data retrieval process.
- System slowdowns or failures due to high demand during peak usage times.

7.4 Predicted Issues

- **Storage:**
    - Handling large volumes of time-series data could lead to memory or storage issues, so managing data efficiently may be a challenge.
    - In an effort to avoid repeating computationally expensive processes, the goal is to store processed data in our system's database to be accessed in the future, but due to the size of the dataset and the user's ability to customise processes, this may be infeasible.
- **Algorithm Optimization:**
    - Ensuring efficient performance of data preprocessing and analysis algorithms on large datasets.
- **User interface:**
    - Since the platform will be used by researchers of varying technical expertise, designing an intuitive user interface that caters to both basic and advanced users may pose challenges.